CS5590-0025 Thomas Jones
Homework #3 08206472

1. Conceptualize a multiple linear regression model considering “mpg (miles per gallon)” as a dependent variable and a set of independent variables, including binary variables for the model year. Write the regression equation in the document.

Before developing a model it is useful to get a gauge of the basic data statistics and metadata.

Feature	Type	Min	Max	Average	Median	Variance
mpg	numeric	9	46.6	23.8	23.0	63.9
cylinders	integer	3	8	5.4	4.0	2.9
displacement	numeric	68	455	191.5	144.0	10915.0
hp	integer	46	230	103.5	92.0	1476.2
weight	integer	1613	5140	2952.5	2750.0	722255.2
acceleration	numeric	8.0	24.8	15.6	15.5	7.9
modelyr	category	70	82	76.0	76.0	13.8
origin	category	1	3	1.6	1.0	0.7
name	character	NA	NA	NA	NA	NA
foreign	category	0	1	0.4	0.0	0.2

Mpg is the dependent variable, and name is character so those two features will not be part of any model data.

When considering what features to select for the model there the first consideration is that features that are highly correlated should be removed. Foreign and origin meet this criteria as any car with an origin of Europe or Asia will definitely have a value of 1 for foreign. Filtering on origin == 1 and foreign == 1 returns no rows, in other words, we can deduce that foreign is a function or origin so we can eliminate foreign from our model.

The dataset includes modelyr already built in as a dummy variable. Since this is a time based criteria the modelyr dummy variable data should be used. This accounts for cases where a particular year, due to legislation changes, for example, may have cause mpg changes not related to the other features included in the dataset, for example, a change in gasoline composition.

Origin should also be converted to dummy variables. As with the modelyr dummy variables the dummy variable trap will be avoided by dropping the first category. For origin, Asia will be dropped. For modelyr, 70 will be dropped.

The cylinders feature could, for this model, be converted into dummy variables since the number of cylinders in a car is a discrete set of limited values, typically 3 to 16 in most modern automobiles. However, unlike the model year or origin, number of cylinders is a characteristic of the vehicle not about the vehicle, i.e. two identical vehicles could be made in different years and in different locations and all other things being equal, inspection of the vehicles would not indicate either of those features. The number of cylinders can be determined by inspection.

The complete equation will then be:

$$\begin{aligned}\text{mpg} = {} & \beta_0 + \beta_1(\text{cylinders}) + \beta_2(\text{displacement}) + \beta_3(\text{hp}) + \beta_4(\text{weight}) + \beta_5(\text{acceleration}) + \\ & \beta_6(\text{origin\_NorthAmerica}) + \beta_7(\text{origin\_Europe}) + \beta_8(\text{modelyr\_71}) + \beta_9(\text{modelyr\_72}) + \\ & \beta_{10}(\text{modelyr\_73}) + \beta_{11}(\text{modelyr\_74}) + \beta_{12}(\text{modelyr\_75}) + \beta_{13}(\text{modelyr\_76}) + \\ & \beta_{14}(\text{modelyr\_77}) + \beta_{15}(\text{modelyr\_78}) + \beta_{16}(\text{modelyr\_79}) + \beta_{17}(\text{modelyr\_80}) + \\ & \beta_{18}(\text{modelyr\_81}) + \beta_{19}(\text{modelyr\_82})\end{aligned}$$

2. Present summary statistics (min, max, mean, median, standard deviation, first quartile, and third quartile) for the variables in your model in a table and briefly comment on the summaries.

Feature	Type	Min	Max	Mean	Median	Std	Q1	Q3
mpg	numeric	9	46.6	23.8	23.0	8.0	17.5	29.8
cylinders	integer	3	8	5.4	4.0	1.7	4.0	8.0
displacement	numeric	68	455	191.5	144.0	104.5	98.0	260.5
hp	integer	46	230	103.5	92.0	38.4	75.0	122.8
weight	integer	1613	5140	2952.5	2750.0	849.9	2208.8	3581.8
acceleration	numeric	8.0	24.8	15.6	15.5	2.8	13.9	17.2

The presence of different means and medians in some of the features indicates skewed distributions, for example cylinders or displacement.

3. Calculate a pairwise correlation matrix for your model variables (while calculating the correlation matrix, only include the dependent variable and the non-binary independent variables that you selected for the model). Present the results in a table and comment on the correlation matrix.

	mpg	cylinders	displacement	hp	weight	acceleration
mpg	1.000	-0.776	-0.803	-0.780	-0.829	0.441
cylinders	-0.776	1.000	0.951	0.845	0.899	-0.506
displacement	-0.803	0.951	1.000	0.899	0.935	-0.550
hp	-0.780	0.845	0.899	1.000	0.867	-0.691
weight	-0.829	0.899	0.935	0.867	1.000	-0.429
acceleration	0.441	-0.506	-0.550	-0.691	-0.429	1.000

The highlighted correlations above show the inter-relationships between cylinders, displacement, hp, and weight. Intuitively, this is to be expected since all are related to the engine of the car, e.g. more cylinders typically means more hp. The model can be simplified by removing as many of these as possible. Displacement has the highest correlation with the other highlighted features. If any correlation > 0.9 is dropped then that means cylinders and weight can be removed from the model.

4. Using R, generate your model and follow these steps:

- **Calculate VIF for the independent variables of your model and show the values in a table.**
- **Comment on the multicollinearity issues in your model.**
- **Avoid multicollinearity issues in your model and re-run the model if necessary.**
- **Show the final model results in a table.**
- **Interpret the coefficients in terms of the sign, size (magnitude), and significance.**

Feature	VIF	Correlation - MPG
cylinders	11.39	-0.776
displacement	24.41	-0.803
hp	11.50	-0.780
weight	12.81	-0.829
acceleration	2.64	0.441
origin_na	2.67	
origin_eu	1.64	
modyr71	2.12	
modyr72	2.08	
modyr73	2.31	
modyr74	2.22	
modyr75	2.39	
modyr76	2.50	
modyr77	2.15	
modyr78	2.45	
modyr79	2.35	
modyr80	2.50	
modyr81	2.38	
modyr82	2.62	

VIF values are shown above for each feature. As discussed in the previous question, cylinders, and weight can be removed from the model as these show VIF > 10.

Feature	VIF	VIF(old)	Correlation - MPG
displacement	8.41	24.41	-0.803
hp	8.70	11.50	-0.780
acceleration	2.06	2.64	0.441
origin_na	2.60	2.67	
origin_eu	1.57	1.64	
modyr71	2.02	2.12	
modyr72	1.92	2.08	
modyr73	2.19	2.31	
modyr74	1.99	2.22	
modyr75	2.08	2.39	
modyr76	2.22	2.50	
modyr77	1.96	2.15	
modyr78	2.24	2.45	
modyr79	2.06	2.35	
modyr80	2.26	2.50	
modyr81	2.15	2.38	
modyr82	2.39	2.62	

Removing those two features reduces the VIF across all the features. Displacement and hp are now below 10.

The estimated parameters are shown in the following table.

Feature	Estimate	Std Error	t value	Pr(>|t|)	α
Intercept	42.116474	2.151777	19.573	2e-16	0.001
displacement	-0.018967	0.004768	-3.978	8.31e-05	0.001
hp	-0.085790	0.013192	-6.503	2.44e-10	0.001
acceleration	-0.302003	0.087756	-3.441	0.000642	0.001
origin_na	-3.089755	0.567114	-5.448	9.08e-08	0.001
origin_eu	-0.885826	0.567444	-1.561	0.119324	
modyr71	-0.616301	0.943701	-0.653	0.514101	
modyr72	-2.187086	0.922111	-2.372	0.018190	0.05
modyr73	-2.233352	0.849690	-2.628	0.008920	0.01
modyr74	-1.206165	0.970289	-1.243	0.214586	
modyr75	-2.196787	0.944589	-2.326	0.020555	0.05
modyr76	-1.078588	0.909418	-1.186	0.236344	
modyr77	0.677214	0.945962	0.716	0.474486	
modyr78	0.679020	0.902085	0.753	0.452075	
modyr79	2.927953	0.954328	3.068	0.002306	0.01
modyr80	7.387880	0.983312	7.513	4.05e-13	0.001
modyr81	4.071293	0.990896	4.109	4.86e-05	0.001
modyr82	6.476725	0.955166	6.781	4.50e-11	0.001

The coefficients indicate that displacement, hp, and acceleration all have a negative effect on mpg. For example, each additional hp reduces mpg by 0.085. Cars from North America get less mpg than cars from other areas, specifically, being from North America reduces mpg by 3.

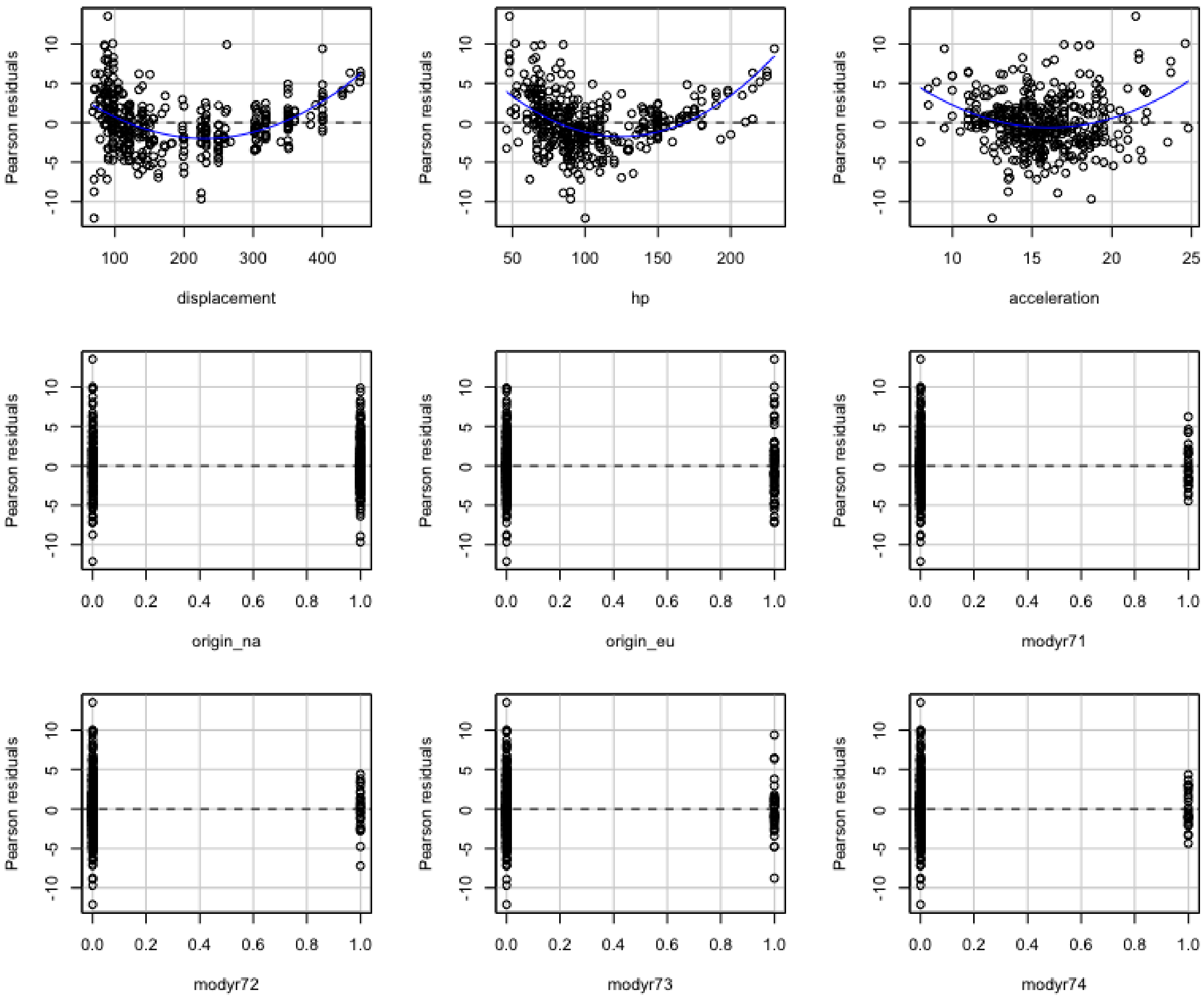
The date of manufacture also has a clear role. Cars built in 1980 get 7 more miles per gallon than cars made in 1971. There is also an obvious general trend increasing mpg from the early 1970’s to the early 1980’s.

The adjusted R^2 for the model is **0.814** with an F-Statistic of 104.7 with a p-value < 2.2e-16 indicating that the features included in the model do have an effect on mpg.

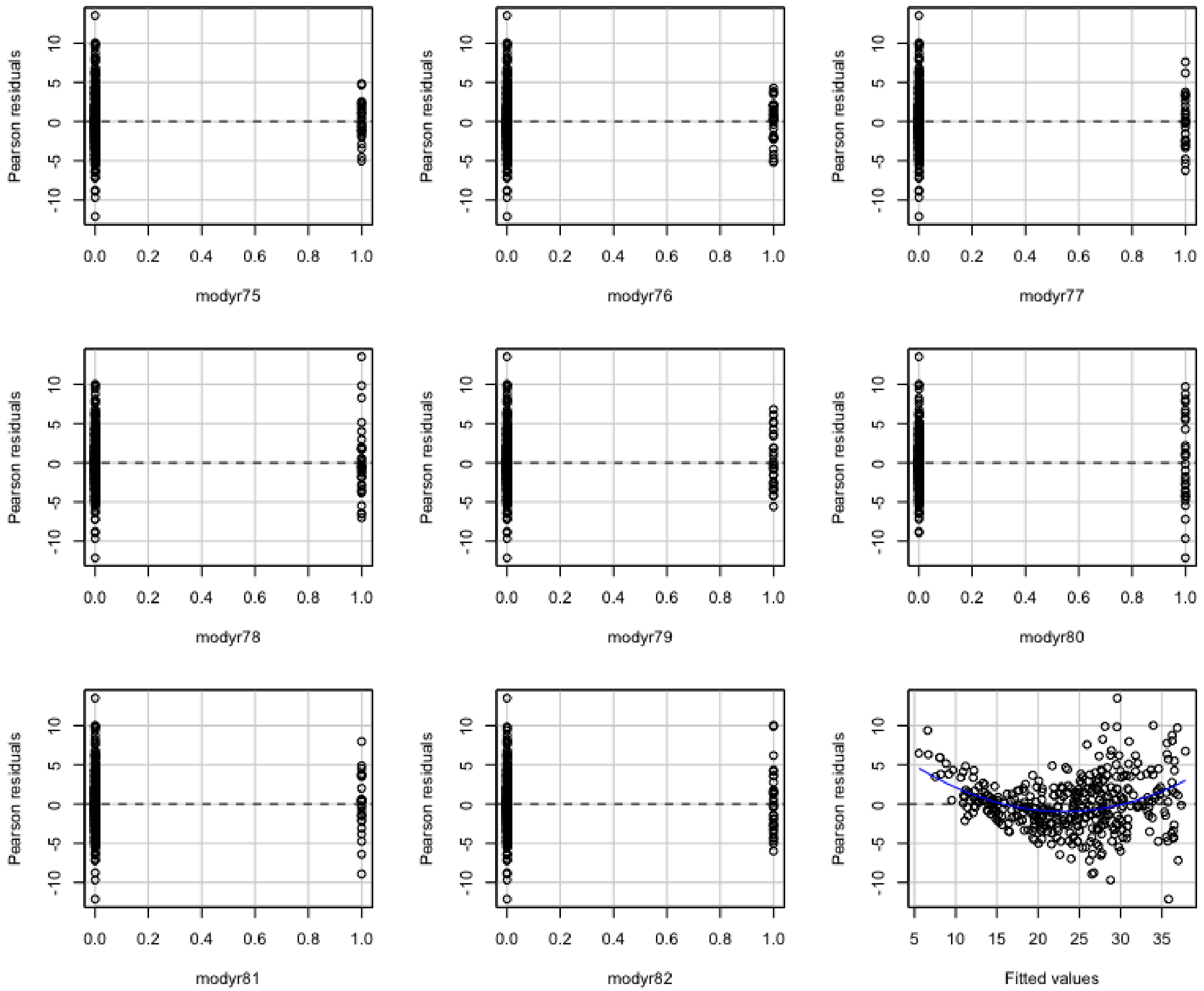
5. Check the following assumptions of the linear regression model:

- **Linearity**
- **Normality in errors**
- **Homoskedasticity**
- **Unusual and influential observations**

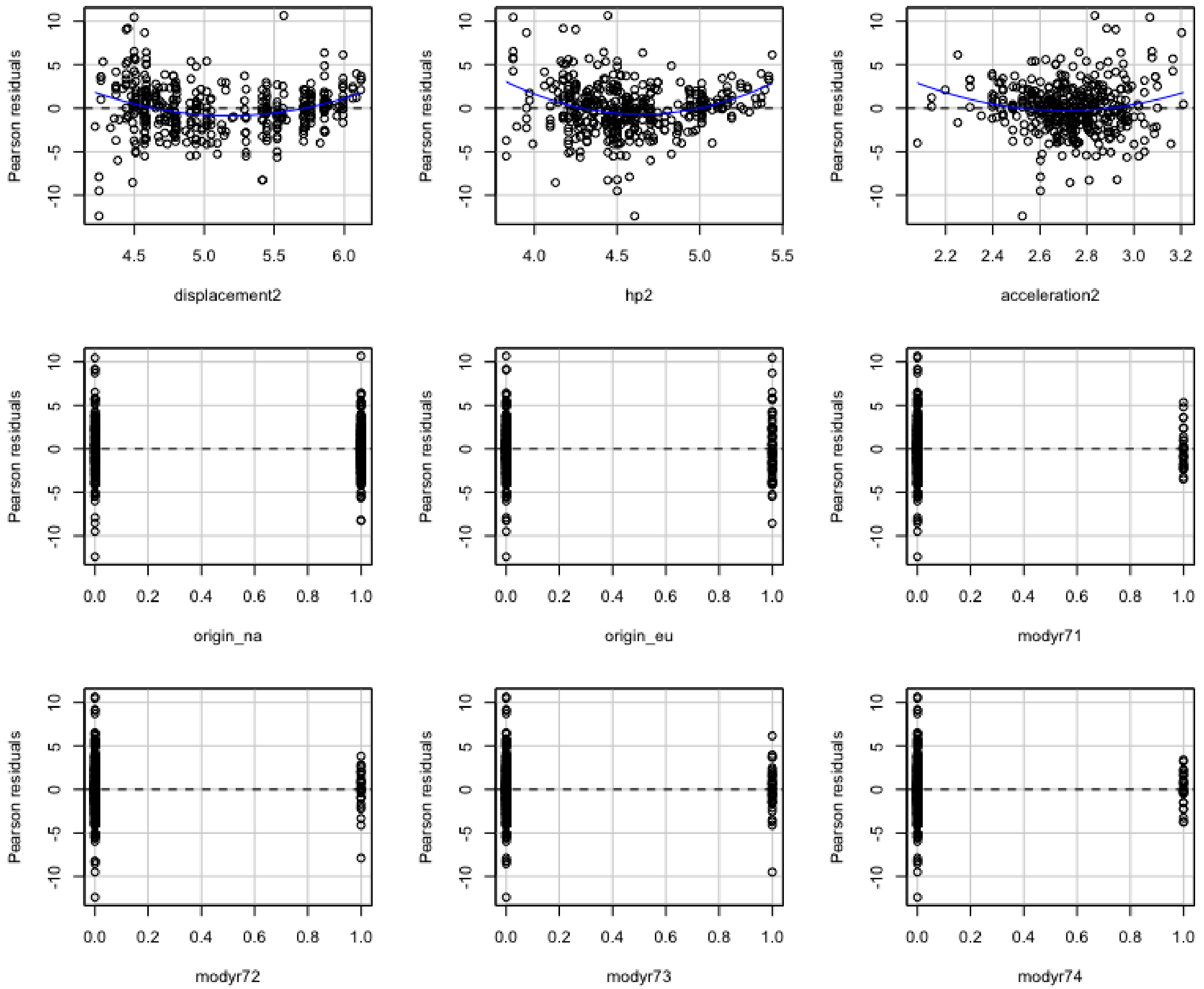
Below are the residualPlots:



The non-linearity of displacement, hp, and acceleration is clear in the residual plots.

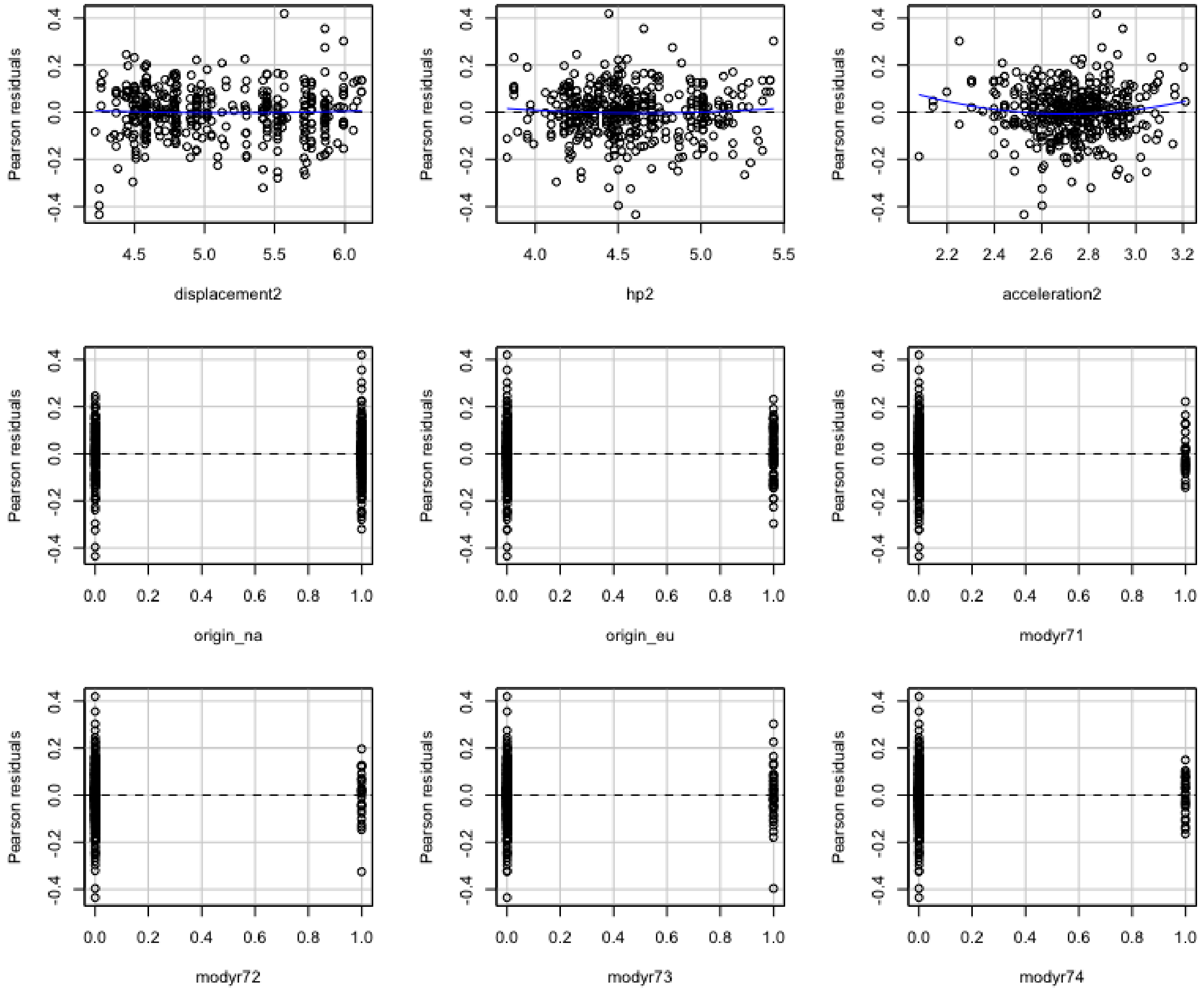


This is still obvious in the fitted values as well. The fit is 104.73.

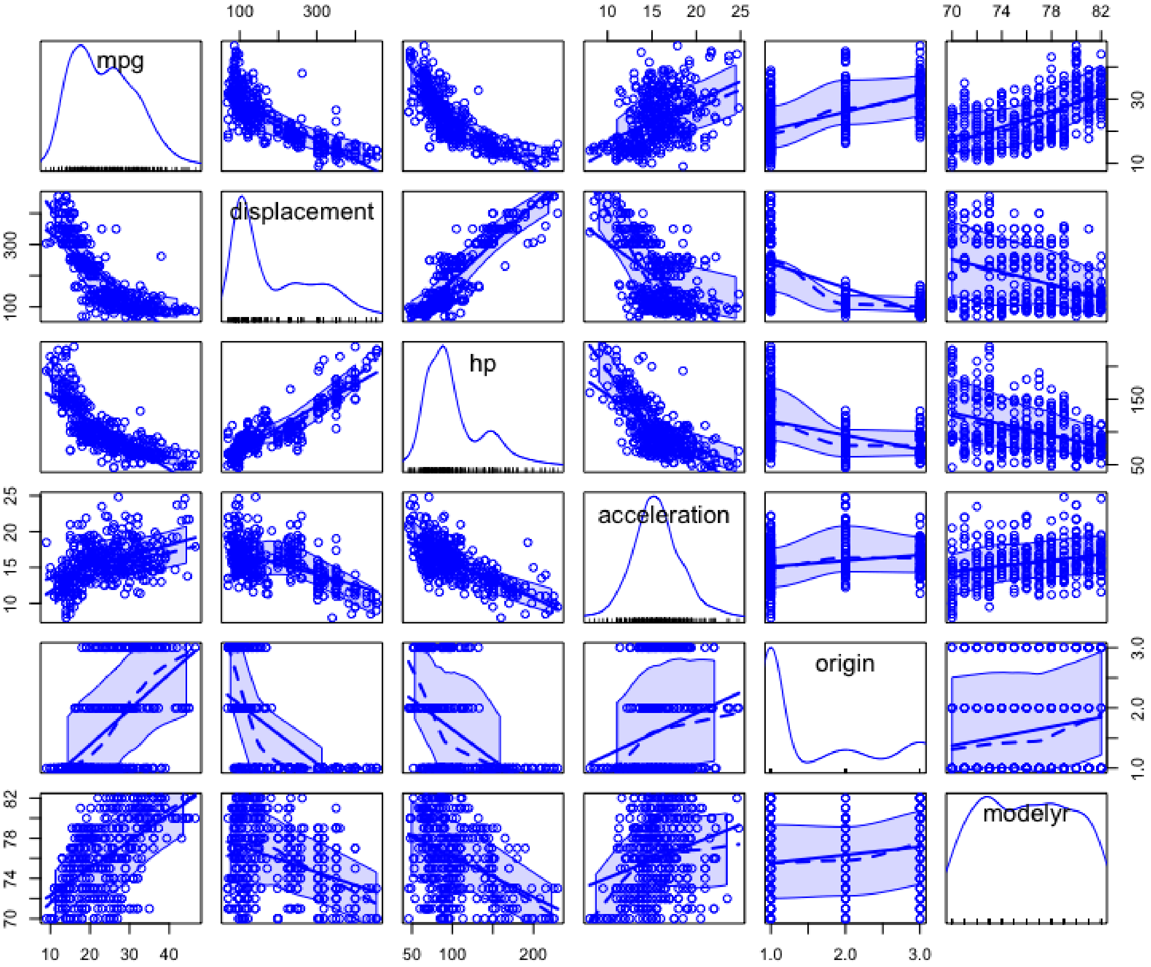


The above residual plots show the effect of using the log of displacement, hp, and acceleration. The curves are still present, but much less pronounced. The calculated fit value is now 155.81 with an improved adjusted R^2 of **0.8672** vs **0.814** from the un-transformed values.

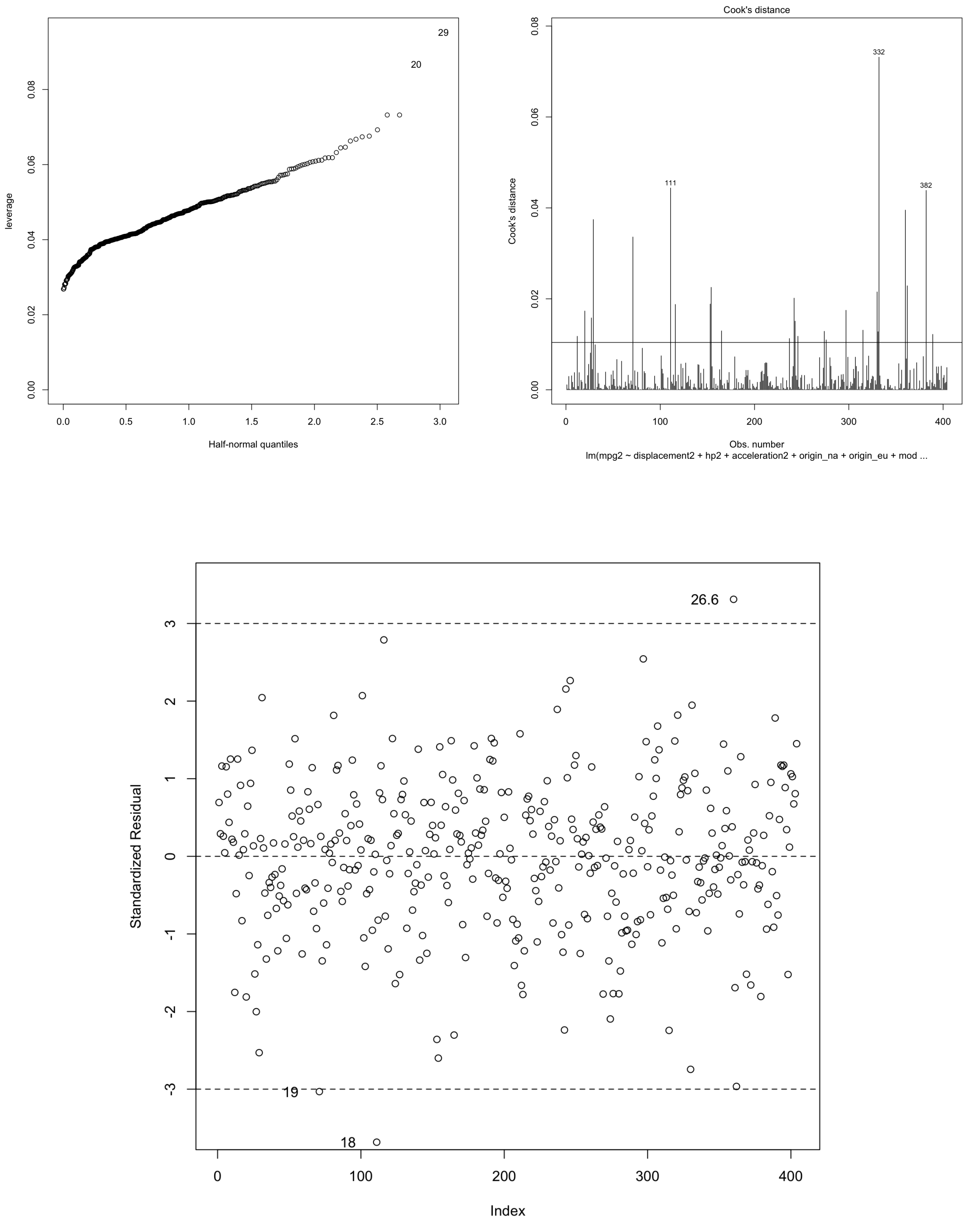
The residual plots also imply heteroskedasticity. Taking the log of mpg as the input and re-running the model yields the following plot:



Displacement and hp are now clearly linear. There is still a non-linearity in acceleration indicated by the plot but this might be an artifact of outliers.

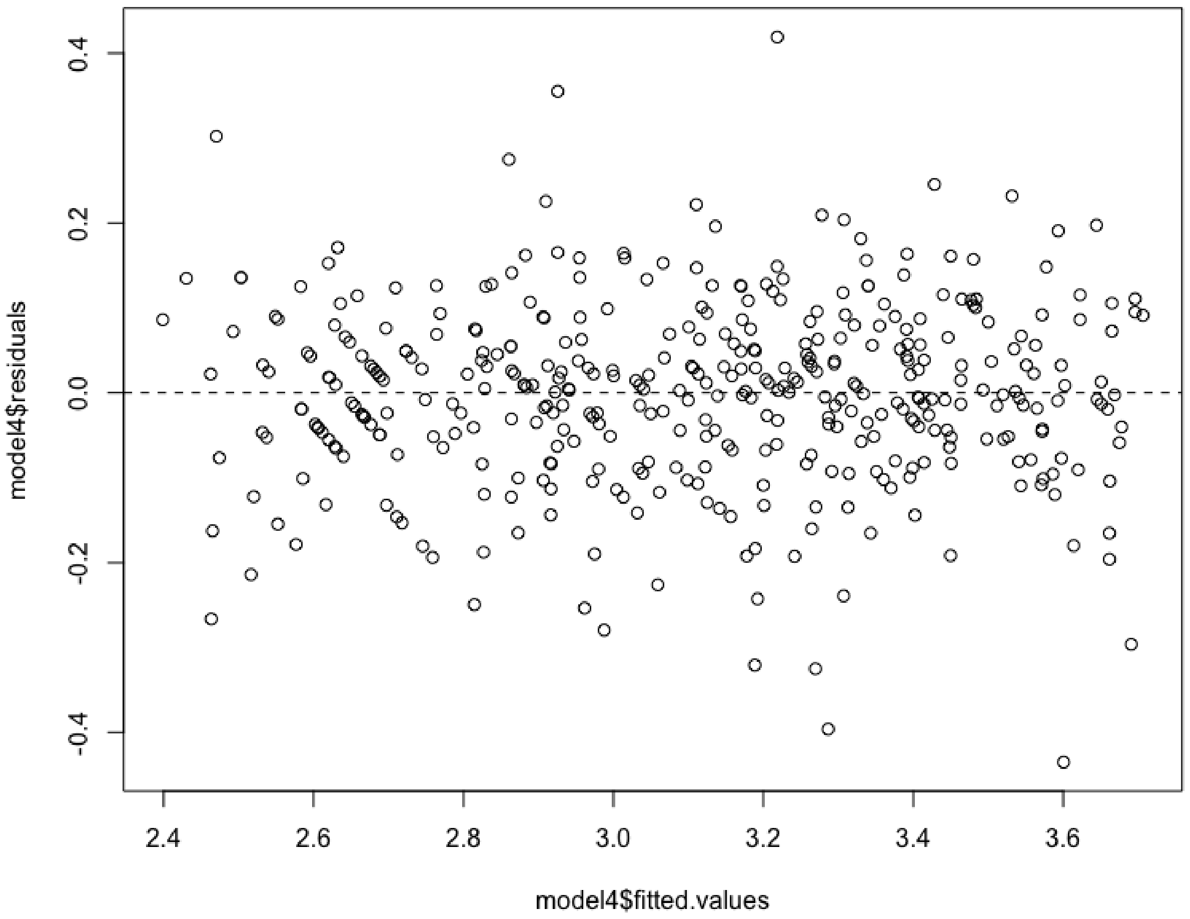


Note, in the scatterplot above, origin and model year are not dummy variables but are the original values. The values are reasonably clustered together with only a few showing as outside the trends.

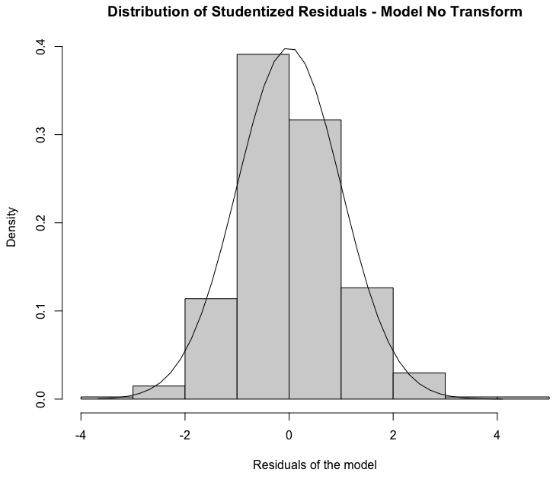


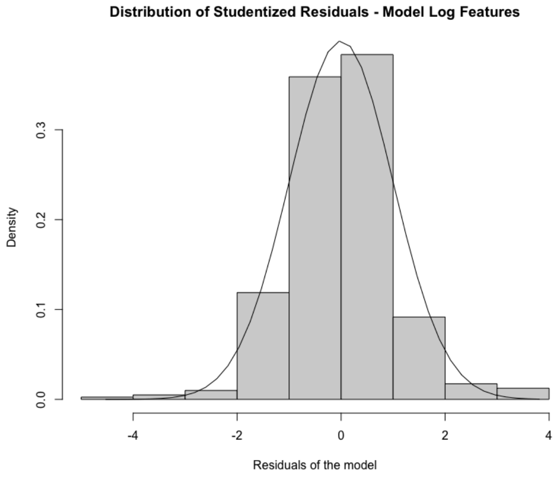
Outlier plots

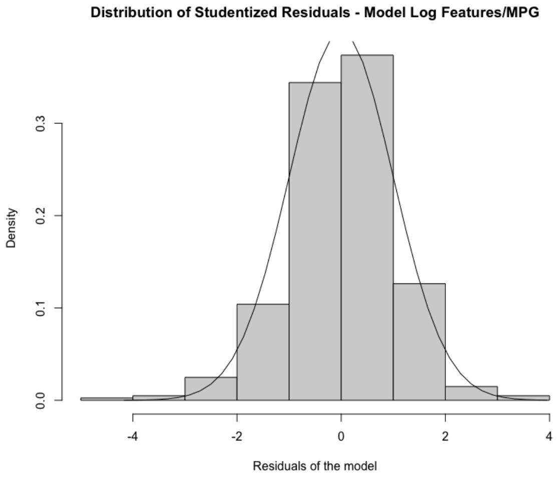
Taking the plot of residuals vs fitted values it can be seen that for the log based dependent and independent variable models heteroskedasticity is not an issue.

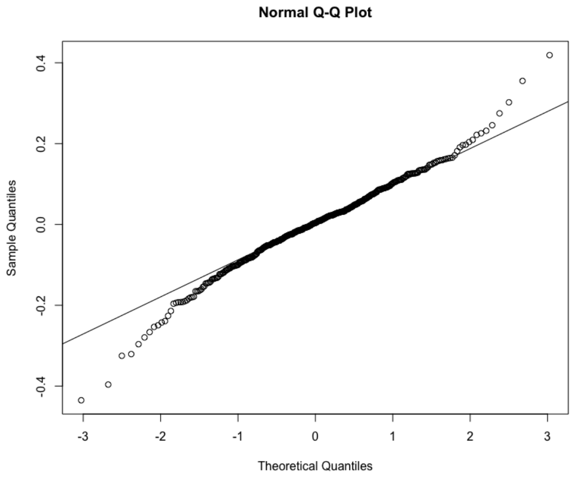


The next step is to check the normality of the errors:









The errors follow normal distributions though the errors without transforming the data show a slight right-skew. Transforming the features shifts the distribution to be better centered. Transforming mpg adds a slight left-skew to the distribution, i.e. favoring an over-estimation of mpg in terms of error.

The Q-Q plot of the residuals shows linear behavior though falling off at the lower and upper quantiles.

The final coefficients for the log fit model are shown below:

Feature	Estimate	Std Error	t value	Pr(>|t|)	α
Intercept	7.604145	0.233527	32.562	2e-16	0.001
log(displacement)	-0.243258	0.28373	-8.754	2.45e-16	0.001
log(hp)	-0.510749	0.044772	-11.408	2e-16	0.001
log(acceleration)	-0.351487	0.046366	-7.581	2.58e-13	0.001
origin_na	-0.036392	0.019534	-1.863	0.063217	0.1
origin_eu	-0.024408	0.018261	-1.337	0.182126	
modyr71	0.006540	0.029983	0.218	0.827443	
modyr72	-0.049888	0.029793	-1.675	0.094842	0.1
modyr73	-0.071042	0.027272	-2.605	0.009543	0.01
modyr74	-0.003717	0.030969	-0.120	0.904529	
modyr75	-0.008675	0.030022	-0.289	0.772778	
modyr76	0.018092	0.028938	0.625	0.532216	
modyr77	0.084577	0.030235	2.797	0.005411	0.01
modyr78	0.102127	0.028767	3.550	0.000433	0.001
modyr79	0.187903	0.030244	6.213	1.35e-09	0.001
modyr80	0.269678	0.031437	8.578	2.37e-16	0.001
modyr81	0.194940	0.031457	6.197	1.48r-09	0.001
modyr82	0.270339	0.030301	8.922	2e-16	0.001

Where adjusted R^2 is **0.8968** with an F statistic of 206.9 for a p-value of 2.2e-16.

Giving a final equation of:

```
mpg = exp(
        7.604 - 0.2432 * log(displacement) - 0.5107 * log(hp) - 0.3514 * log(acceleration)
        - 0.0364 * origin_na - 0.0244 * origin_eu + 0.0065 * modyr71 - 0.0499 * modyr72
        - 0.0710 * modyr73 - 0.0037 * modyr74 - 0.0087 * modyr75 + 0.0181 * modyr76
        + 0.0846 * modyr77 + 0.1021 * modyr78 + 0.1879 * modyr79 + 0.2697 * modyr80
        + 0.1949 * modyr81 + 0.2703 * modyr81
        )
```

A sanity check using the averages of displacement, hp, and acceleration with a North America origin and a model year of 1976 yields.

$$mpg \ = \ e^{7.604 - 0.2432\, ln(191.5) - 0.5107\, ln(103.5) - 0.3514 \ \ ln(15.6) - 0.0364 + 0.0181}$$

= 19.54, compared to a mean of 23.8 for the dataset.

6. Are American cars heavier than foreign cars? Use an appropriate statistical test (make sure to write each step and the relevant values in detail in the document).

For this test two population means will be compared. The cars for America will be considered a different population than the foreign cars so a two population means test will be used where the variances are not known.

The first step is to state the hypothesis, in this case H_0, that American cars are not heavier than foreign cars. The alternative hypothesis, H_a, is that American cars are heavier. In other words,

$$\bar{x}_{\text{American}} > \bar{x}_{\text{foreign}}$$

$$H_0 = \bar{x}_{\text{American}} \leq \bar{x}_{\text{foreign}}$$
$$H_a = \bar{x}_{\text{American}} > \bar{x}_{\text{foreign}}$$

These can be tested in R using t.test.

```
t.test(df.american$weight, df.foreign$weight, conf.level = 0.99, alternative = "greater")
```

Where df.american and df.foreign are the filtered car data based on the foreign field.

The results from the test give a t = 17.569 indicating $\bar{x}_{\text{American}} > \bar{x}_{\text{foreign}}$ with a p-value of 2.2e-16 which is less than the α of 0.01 indicating the null hypothesis is rejected, thus we accept the alternative hypothesis that American cars are heavier than foreign cars.

7. Do the model year binary variables jointly have explanatory power? Use an appropriate statistical test (make sure to write each step and the relevant values in detail in the document).

The null hypothesis, H_o, is that the binary values for model year do not have any explanatory power, i.e. do not affect the model outcome.

The alternative hypothesis, H_a, is that the binary values for model year do explain changes in mpg.

The car package in R provides the linearHypothesis method. Given the original model and a set of columns as a hypothesis, in this case the binary columns for modyr71 to modyr82, the function returns a F statistic and the p-value.

For the model using log(mpg), log(displacement), log(acceleration), origin_eu, and origin_eu, including modlyr binary values the resulting p-value is 2.2e-16, therefore H_o is rejected and model year binary variables do have explanatory power over the model.

8. Calculate model's goodness of fit and comment.

The calculated adjusted R^2 of the log based model is 0.8968 (based on the summary method from R). This indicates that 89% of the data variance is explained by the model.

The usability of the model depends on the accuracy necessary.

Note, a final test was performed using both the log and non-log versions of the variable and including cylinders as well as weight, i.e.

```
model5 = lm(formula =
        mpg2 ~ cylinders + weight + log(cylinders) + log(weight) +
        displacement2 + displacement + hp2 + hp + acceleration2 +
        acceleration + origin_na + origin_eu +
        modyr71 + modyr72 + modyr73+ modyr74 + modyr75 +
        modyr76 + modyr77 + modyr78 + modyr79 + modyr80 +
        modyr81 + modyr82 ,data = df_mod)
```

The final adjusted R^2 for this model was 0.9087, i.e. a full percentage point gain but at the expense of additional computational complexity. One interesting observation is that the coefficients for the non-log features took on opposite signs of the log features (except hp) which hints at a more complex relationship between mpg and those raw values.

Fuller model coefficients (output from R-studio):

```
Coefficients:
                 Estimate     Std. Error   t value     Pr(>|t|)
(Intercept)      1.039e+01    1.792e+00    5.800       1.40e-08 ***
cylinders        -1.143e-01   5.277e-02    -2.165      0.03098 *
log(cylinders)   5.727e-01    2.934e-01    1.952       0.05169 .
weight           2.685e-05    9.081e-05    0.296       0.76764
log(weight)      -6.068e-01   2.923e-01    -2.076      0.03860 *
displacement2    -1.878e-01   1.254e-01    -1.497      0.13516
displacement     7.833e-04    5.960e-04    1.314       0.18959
hp2              -9.818e-02   1.465e-01    -0.670      0.50312
hp               -1.484e-03   1.282e-03    -1.158      0.24776
acceleration2    -8.177e-01   2.999e-01    -2.726      0.00670 **
acceleration     4.334e-02    1.887e-02    2.297       0.02215 *
origin_na        -3.215e-02   1.944e-02    -1.654      0.09901 .
origin_eu        3.984e-03    1.833e-02    0.217       0.82802
modyr71          3.076e-02    3.077e-02    1.000       0.31811
modyr72          3.646e-03    2.991e-02    0.122       0.90306
modyr73          -3.278e-02   2.708e-02    -1.211      0.22679
modyr74          6.128e-02    3.203e-02    1.913       0.05650 .
modyr75          5.742e-02    3.102e-02    1.851       0.06494 .
modyr76          8.195e-02    2.955e-02    2.774       0.00582 **
modyr77          1.446e-01    3.066e-02    4.715       3.41e-06 ***
modyr78          1.562e-01    2.894e-02    5.397       1.19e-07 ***
modyr79          2.457e-01    3.052e-02    8.051       1.07e-14 ***
modyr80          3.485e-01    3.163e-02    11.020      < 2e-16 ***
modyr81          2.685e-01    3.182e-02    8.438       6.92e-16 ***
modyr82          3.304e-01    3.047e-02    10.844      < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1041 on 379 degrees of freedom
Multiple R-squared:  0.9141,  Adjusted R-squared:  0.9087
F-statistic: 168.1 on 24 and 379 DF,  p-value: < 2.2e-16
```